Whole Genome Analysis: techniques and pitfalls

Todd Skaar, Ph.D.
Indiana University

March 31, 2008

Types of genetic variations

Single nucleotide polymorphisms (SNPs): Single base pair changes in the genome in a population.

Copy number variations (CNVs): Interindividual variations in the number of copies of a specific gene or chromosomal region.

Insertions and deletions (Indels): Regions of DNA that are either inserted into or deleted from the genome.

Allelic frequency: The number of variant alleles divided by the total number of alleles in a pop’n

Terminology Definitions: Genome Wide

Geographic SNPs: These are SNPs chosen in attempt to get SNPs roughly equally spaced across the genome.

HapMap SNPs: SNPs selected based on the haplotyping data

Genome-wide SNPs: SNPs that are located across the whole genome; this does not mean “every SNP” in the genome.

These are “discovery-based”, rather than “hypothesis-based”.

Terminology Definitions: Candidate genes

Candidate genes: These are genes for which there is a biological reason as to why they may be involved in the response to the drug of interest.

Functional SNPs: These are SNPs that are known to alter the function of a gene or gene product.

These are "hypothesis-based", which are based on knowledge of the genes and drugs.

Gene resequencing

What is it? The identification of genetic variants by DNA sequencing; it is usually focused on a specific gene and within in a population.

When do it? When the genetic variants within a gene are not well characterized; this usually due to limited sample number in previous studies

Why do it? To discover new SNPs, their frequencies, their ethnic distributions, and their haplotype structures.

How does it relate to GWAS? Needed to identify SNPs to put on chips; maybe used to identify functional SNPs after GWAS.

General properties of genotyping platforms

	SNP discovery	Initial cost	Cost/ SNP	Flexibility	Sample Throughput	SNP Throughput
Sequencing	yes	low	high	high	low	low
Taqman	no	low	mid	high	high	low
Sequenome	no	low	mid	mid	high	mid
Luminex	no	mid	high	mid	low	mid
GWAS	no	high	low	low	low	high
NextGen	yes	high	low	low	low	high

Red: Good
Pink: Intermediate
White: Poor

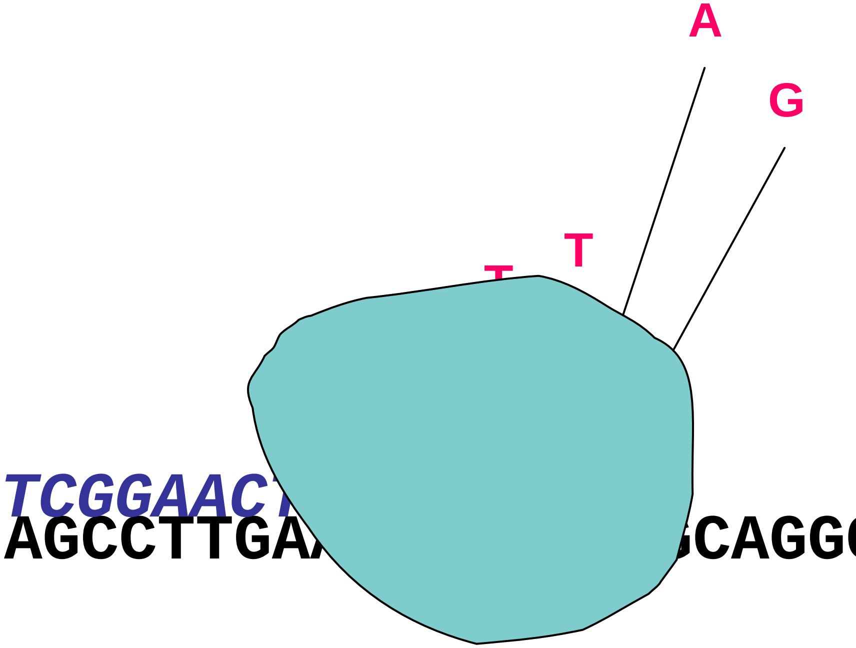
Traditional DNA gene resequencing

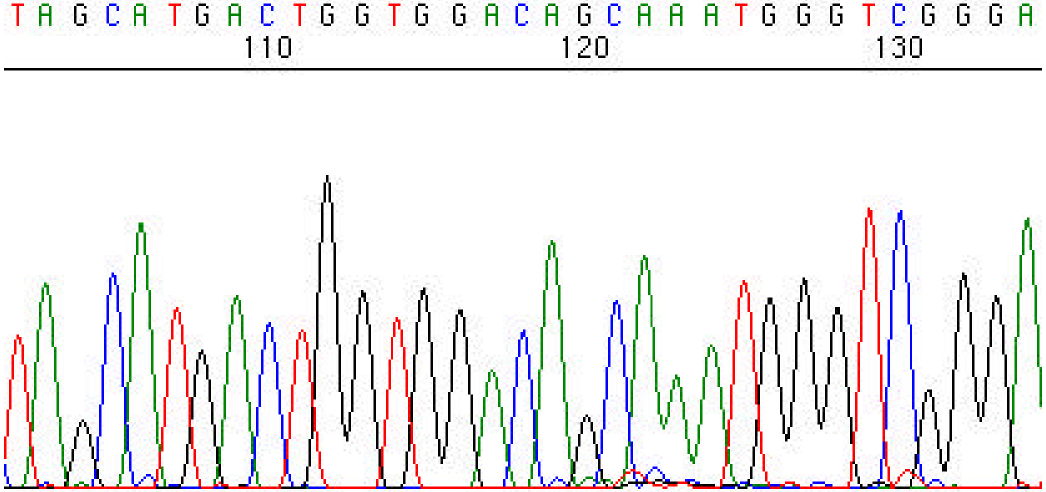
- Region of interest amplified by PCR
- PCR products labeled with fluorescent tags
- Fluorescent products analyzed on capillary electrophoresis “sequencer”
- Reactions done in 96-well plates

Traditional DNA gene resequencing

- Generally regarded as the gold standard for accuracy
- Usually used for SNP discovery
- Readily available
- Low throughput
- Rarely used for routine genotyping
- Rarely multiplexed (multiple variants in one tube)

Traditional DNA sequencing





Taqman assays

PCR based

Allelic discrimination by allele specific probes

96/384-well format

Single sample/SNP per well

Analyzed in real-time PCR thermocycler

Completed in a single PCR reaction step

Arrays becoming available (33 nl; 3072/chip)

Taqman assays

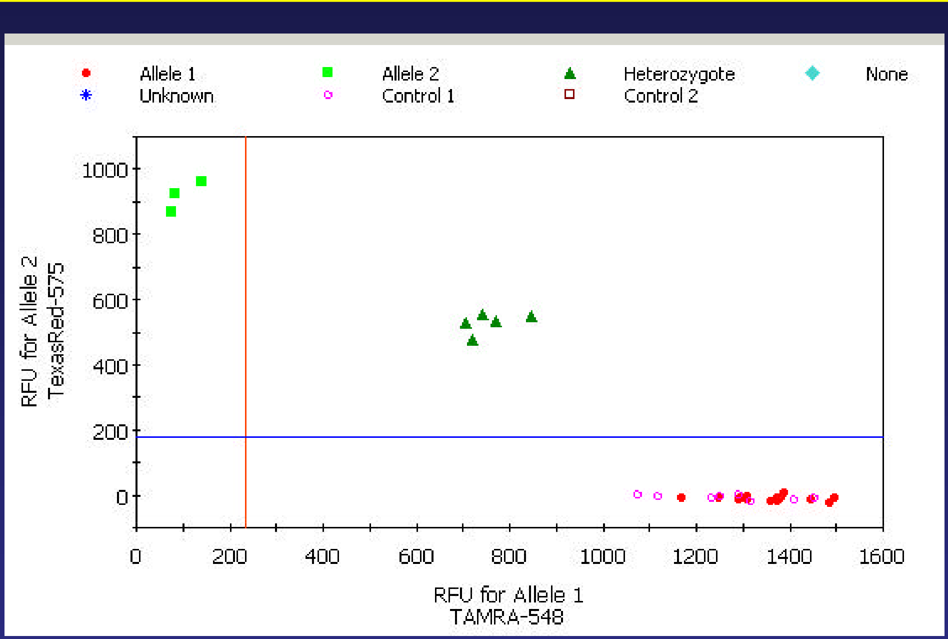
Pro's:

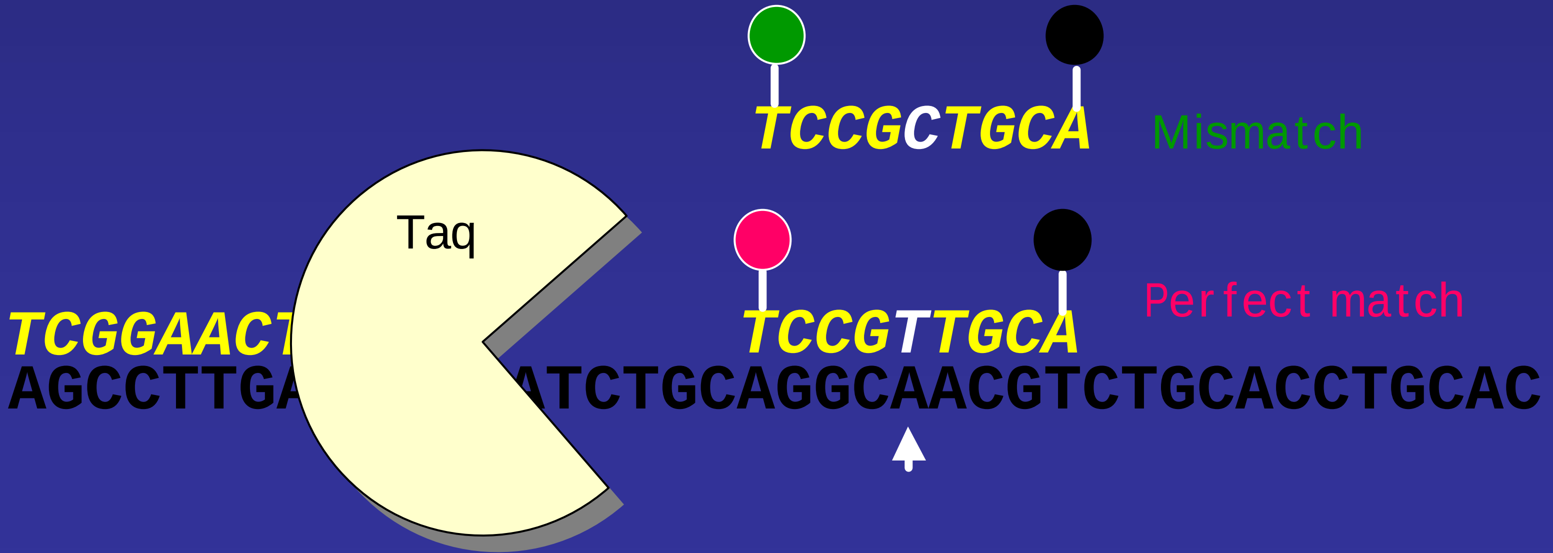
- instrumentations readily available
- many assays already designed
- DME assays available
- easy to add more
- fast run time
- low startup cost
  ($100-300/SNP = ~300 samples)
- arrays platforms becoming available

Cons:

- no multiplexing (i.e. one snp per tube)
- cost per snp is expensive (~$1/sample/snp)
  unless using new arrays

Taqman assay





TCCGCTGCA

Mismatch

Taq

TCCGTTGCA

Perfect match

TCGGAACT

AGCCTTGA ATCTGCAGGCAACGTCTGCACCTGCAC

GWAS: Illumina, Affymetrix Chip

Affymetrix:

- allelic discrimination by hybridization
- oligos synthesized on silicon chips
- genomic DNA labelled with fluorescent tags
- hybridized to chips
- read in chip reader

Illumina:

- allelic discrimination by oligonucleotide ligation
- thousands of oligos synthesized together
- oligos flanking SNPs are hybridized to DNA
- oligos ligated together
- analyzed on fiberoptic chips

GWAS: Illumina, Affymetrix Chip

Pro’s:

- Highly multiplexed
  - WGA: 300,000 – 1.5 million/array
  - custom chips are often 1536 SNPs
- low cost per SNP (~$0.002/SNP)
- customizable
- most places have the instrument

Con’s:

- expensive startup costs
  ($50,000 for ~500 samples)
- not easy to modify after first order
- instrument is expensive
- difficult for highly homologous genes (P450’s)
- typically only SNPs with frequent allele freq’s

Why do GWAS studies miss some genotype

1. Gene has no SNPs on the chip

2. Gene has snps on the chip, but not the is there, but not the right ones

   - rare snps usually not included
   - may not be enough within a gene to get all

3. many are intronic snps

Illumina GWAS

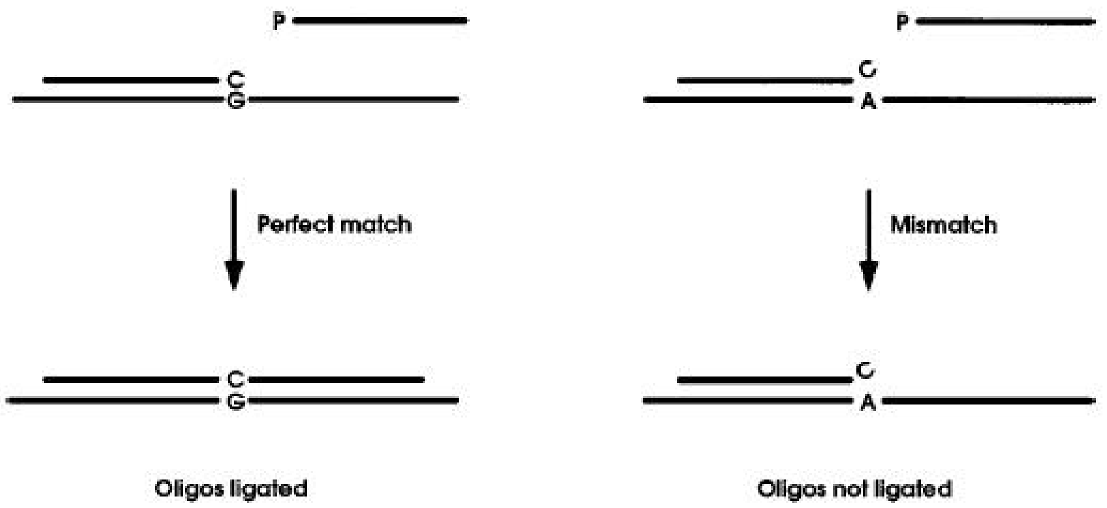


Figure 4 Allele-specific oligonucleotide ligation.

Next Generation DNA Sequencing

- Very early in applications
- Allelic discrimination by sequencing
- Thousands of individual mini-sequencing reactions on a single plate
- Get millions of base pairs of sequence per run
- Currently one run per sample maybe able to combine samples
- Sequence capture arrays becoming available to focus sequencing on genes of interest

Next Generation DNA Sequencing

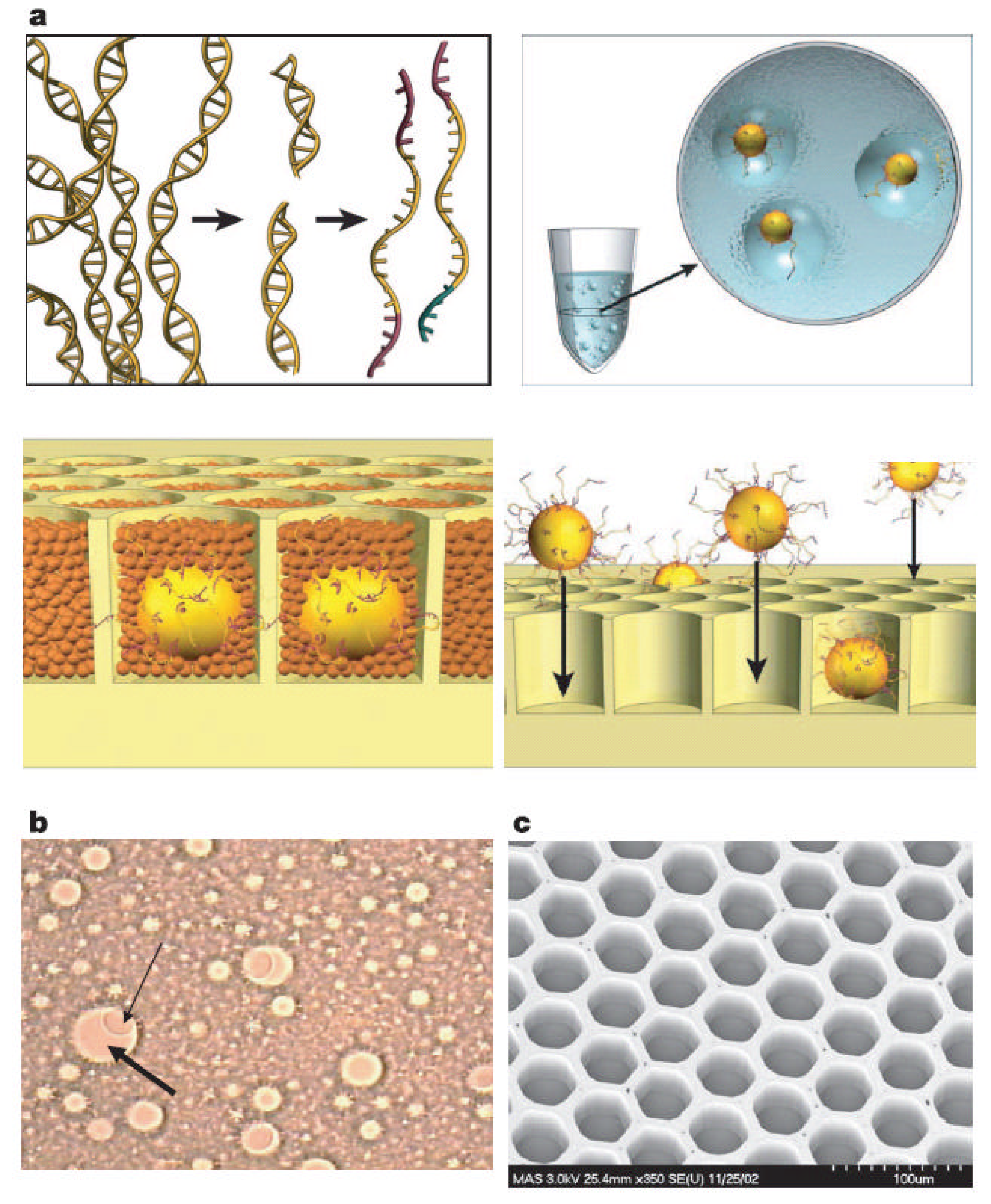
Pro:

- Comprehensive analysis of each gene in full
- Works for SNP discovery

Con:

- Expensive instrument
- Expensive reagents
- Low sample throughput
- Early phase of technology development
- Instruments not readily available

Next Generation Sequencing



Margulies et al, 2005
Nature

General properties of genotyping platforms

	SNP discovery	Initial cost	Cost/ SNP	Flexibility	Sample Throughput	SNP Throughput
Sequencing	yes	low	high	high	low	low
Taqman	no	low	mid	high	high	low
Sequenome	no	low	mid	mid	high	mid
Luminex	no	mid	high	mid	low	mid
GWAS	no	high	low	low	low	high
NextGen	yes	high	low	low	low	high

Red: Good
Pink: Intermediate
White: Poor

Title

Text